International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Efficient Load Balancing in Heterogeneous Cloud to Reduce Response Time and Processing Time Using BUSY & AVAILABLE Algorithm

Matsvimbo Davison, David Fadaraliki

[1]School of Information Science& Technology, Harare Institute of Technology, Harare Zimbabwe
[1]*h190065c[at]hit.ac.zw*
[2]*dfadaraliki[at]hit.ac.zw*

Abstract: ***The unstoppable ever - increasing use of the internet by millions of users across the globe triggers the imperativeness of cloud computing. A greater pool of organizations and developers had paid a magnificent interest to cloud computing. In cloud computing, the advantages obtained outweigh by far legacy on - prem data centre technology. With cloud computing users enjoy the pay - as per use model while appreciating an assortment of facilities from applications, process competence and storage. Henceforth, the drastic movement of organizations toward the cloud is overwhelming. Cloud computing is now miles ahead of conventional computing. In this research paper, a "BUSY & AVAILABLE" Algorithm is proposed for virtual machine allocation in a heterogeneous cloud. The proposed algorithm distributes autonomous user requests to available VMs in datacenter efficiently to manage appropriate load balancing***

Keywords: Load Balancing, Autonomous Task, Heterogeneous Cloud, Response Time, Datacentre Processing Time

1. Background

Industry and academia had shown a great interest in cloud computing at the time of this writing. Compliant to the official NIST definition, "cloud computing is a model for enabling ubiquitous, convenient, on - demand network access to a shared pool of configurable computing resources (for example, networks, servers, storage, applications and services) that can be expeditiously provisioned and released with least possible management effort or service provider synergy. "[1]The ever - increasing gravity of cloud computing in the IT world can't be thwarted. Cloud computing can slash operational and capital costs. Cloud computing is a very colossal concept. One of the outstanding features of cloud computing is its elasticity at which the resources can be scaled up or down based on some tracked metrics of the cloud computing services. Cloud computing seems to be the favorite emerging technology which is drawing the scrutiny of the entire technocrat in the field of information technology. It is necessarily indicated to as accessing computing service over the cyberspace.

1.1 Motivation

The tremendous adoption of the cloud by millions of organizations as a result of its valued features and services has resulted in substantial traffic at the cloud service provider. Load imbalance may be as a result of the dynamic workload patterns and a massive quantity of service request[2]. Improper load balancing has a greater contribution in terms of resources wastage and performance degradation on of the service provider. Henceforth, the end users suffer a degraded Quality of Service (QoS) and violates SLA [3]. Incoming user requests to the datacentre are processed by allocating them to a VM resource. A heterogenous and homogenous resource configuration may be used for a modern datacentre. Allocating user requests and balancing the load amongst available VMs had proved to be a mammoth task on heterogenous VM configuration [4]. A very crucial task in cloud computing is how to efficiently load balance

1.2 Load Balancing

As the influx of traffic escalates on a website or business application, it's unbearable for a single server to support the full workload. Organizations thus must spread the workload over several servers and to accomplish this, load balancer tools can help in dividing the network traffic consistently and thus preventing fiascos caused by overloading a specific resource. Hence, while load balancing and its tools can improve the performance and availability of applications, websites, databases, and other computing resources, it can act as an imperceptible facilitator to guarantee the connection requests are perceptible to end users [5]. High Quality of Service (QoS) is the customer expectation from the service provider as well as profit is awaited by the provider. These expectations can both be achieved through maximizing resource utilization by proper load balancing [1]. While allocating user requests to the VMs, the algorithm should consider current VMs' state and the heterogeneity of resources in cloud data centre.

2. Literature Review

The ever - increasing gravity of cloud computing in the IT world can't be thwarted and its performance is one greatest concern. The major challenge noted in cloud computing is load balancing. In this section research studies related to load balancing are highlighted.

2.1 Ojhaet al in 2014 proposed [6] a hybrid load balancing approach was proposed using "Round Robbin" and "Throttled" algorithm**.** The throttled algorithm selected the

Licensed Under Creative Commons Attribution CC BY

VM allocated using Round Robbin method. The throttled algorithm efficiently uses all the VMs it selected. Throttled algorithm performs great in a homogeneous cloud environment, nevertheless performance might get degraded in case of heterogenous cloud environment.

2.2 Badshaha P Mullaet al in 2020 proposed[1] that the heterogeneous configuration of cloud datacentre resources and the genuine utilization of processing elements have not been accounted for during the load balancing mechanism. For efficient utilization of all the available resources at cloud datacentre reliant on their processing competences these issues are supposed to be considered by the efficient load balancing algorithm. These considerations are imperative inputs to improvement of datacentre processing time and quality of Service (QoS) to the cloud users. The In the previous studies, throttled algorithm is deemed best algorithm over the traditional load balancing algorithm.

2.3 Nilesh et al in 2017 proposed[7] that in 1992, Macro Dorigo and his colleagues proposed Ant Colony Algorithm (ACO). ACO isinspired from Real Ants. When hunting for a food, Ant has aptitude to determine the pathbetween nest and food. When ants explore for a food, ant wanders haphazardly and in return tripthey lay some chemical substances i. e., pheromone on the ground. The path withthe highest density of pheromone will attract all the other ants to follow that same path for food search and back to the nest. The pheromonetrails are used by ants to reach the food sources. The Indirect communication amongst the ants via pheromone trails permits them to get theshortest paths between their nest and food sources.

2.4 Sandeep Sharma et al 2008 proposed [8]Round Robbin is a widely used, straight forward and easy algorithm. A circular fashion is used by the algorithm to assign user requests to each VM, and this is done without taking into account the processing capabilities of any individual VM. RR is extremely efficient for datacentres where all VMs have equal processing capabilities. Put simply, it works perfectly for the homogeneous cloud.

2.5 Bharat Khatavkaet al 2017 proposed [9] that so as to improve continuity and availability of cloud computing, an introduction of a load balancing approach that reduces cloud latency and response time was done. A study was carried out to copy data from a source faulty VM to a target VM, to make sure users seamlessly access information uninterruptedly. the reason behind is to put together weighted RR and Max Min algorithms to come up with an efficient Weighted Max min algorithm and two important parameters that is waiting time and response times were reduced.

3. Methodology

The aim of this research is mainly centred at minimizing the user request response time and heterogenous cloud datacentre processing time. The proposed algorithm which is deployed at the Vmloadbalancer in consultation with the Datacentre Controller (DCC), identify the suitable VM for the user request allocation. “Round Robin”, “Throttled” and “Equally Spread Current Execution Load” algorithms are compared with the proposed algorithm for performance analysis.

3.1 Proposed Algorithm

VMs processing power is considered during the allocation. Two separate tables viz AVAILABLE Table and BUSY Table are used to store the VM indexes of the proposed algorithm [1].

Upon receiving the user request, to find out the VMthe AVAILABLE table is scanned. Incase no VM was found in the AVAILABLE table, the VM with the suitable capacity will be searched by the algorithm from the BUSY tableto execute the current request [10]. To avoid extra overload on the BUSY VM defined threshold levels are used.

The available VM resources in a heterogenous cloud environment may have varying capabilities in terms of memory, processing speed and number of processors. Accordingly, BUSY Table having a VM with resources can be used to execute a request rather than putting it in the waiting queue [1].

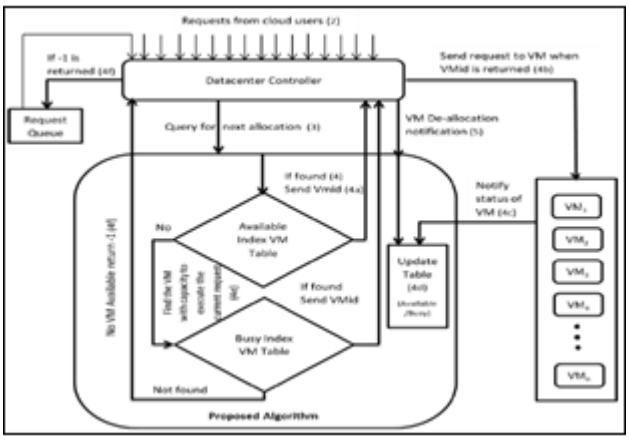

Figure 1: Efficient Load Balancing in Heterogeneous Cloud to Reduce Response Time and Processing Time Using proposed “BUSY & AVAILABLE” Algorithm

The proposed algorithm is supposed to calculate capacity of every VM so that a suitable VM from the BUSY table is searched. The average capacity of the VM is calculated and the only VM having its capacity greater than or equal to average capacity is selected. Round Robin manner is used to allocate the requests. Two (2) threshold values are used by this algorithm so as to get rid of extra overload on BUSY VM.

4. Simulation and Result Argument

A simulator has been used to carry out the experiment as real tests limit experiments due to infrastructure scale[1]. It is time consuming to carry out performance measurement of the system on real cloud environment [11]. The repetition of experiments is very difficult in real cloud [12]. Overall, it is very expensive to access infrastructure of a real cloud.

Licensed Under Creative Commons Attribution CC BY

4.1 Simulation Setup

For the simulation to be carried out, a cloud based social networking application on internet have been considered that is FB. Facebook Subscribers and World Population Statistics updated as of March 31, 2021 across the main seven (7) regions [13].

In this research, a similar system has been assumed on the normalized scale (1/300th). Seven (7) User Bases have been defined which represents the users from the above 7 regions. Table below represents user base characteristics.

Table 1: User base characteristics

User Base & Region	Time Zone (GMT)	Peak Hours	AvPeak Users	Av Off - Peak Users
UB - 0	GMT+200	18.00–20.00	883 000	883 000
UB - 1	GMT+600	01.00 – 03.00	3748000	3748000
UB - 2	GMT+400	20.00 – 22.00	173700	173700
UB - 3	GMT+400	15.00 – 17.00	152800	152800
UB - 4	GMT+600	01.00–03.00	484800	484800
UB - 5	GMT+600	13.00 – 15.00	880500	880500
UB - 6	GMT+1000	09.00 – 11.00	81800	81800

To make life easier here, an assumption of 10% of the users are active during off - peak hours. Another assumption made here is that when a user in online for every five (5) minutes a new request is generated[1]. As many users access the application during the night after working hours for about 2 hours, henceforth thus how peak hour was concluded from.

VMs are hosted on physical machines residing in a datacentre. Different capacity for physical machines such as number of CPU (2/4/6/8/10), RAM (2GB/4GB) and processing power in terms on Millions of Instructions per Second (MIPS) enables creating a heterogeneous cloud environment.

5. Results and Discussion

For analysing the behaviour of the "BUSY and AVAILABLE" algorithm, three different scenarios have been considered here. For each scenario, number of VMs, broker policy and number of data centres are different. A repetition for the simulation on "Throttled", "Round Robbin" and the proposed "BUSY and AVAILABLE" algorithms has been done for analysing results.

Scenario 1: Single Datacentre with 50 VMs

In the figure 5 below, it is clearly shown that "Round Robbin" and "Throttled" takes more time for performance parameters considered. In "Round Robbin" algorithm, allocation of VMs is done in a circular manner not considering its processing capabilities such as RAM, number of processors, processing power and so on. In this case, only one data centre is used and hosted with fifty (50) VMs. "Throttled" algorithm allocates a single VM only a specified number which is represented by threshold value, the user requests at particular any given time. In a scenario where extra requests are there; these requests are put in a waiting queue until the availability of the next VM. Processing capability of VMs is not considered too in "Throttled" algorithm. The proposed "AVAILABLE&BUSY" Algorithm has considered the capabilities of each VM and request were allocated accordingly.

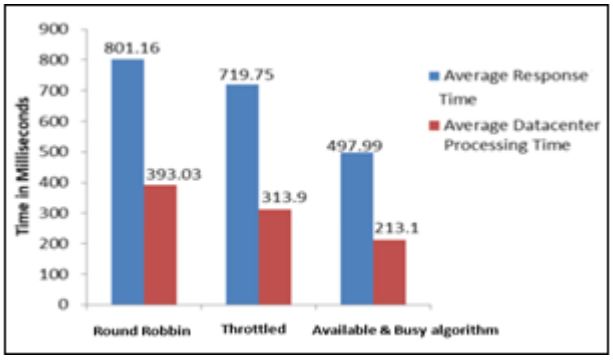

Figure 2: Single Datacentre with 50 VMs

In the proposed "Busy and Availability" algorithm; processing time of data centre reduced by 32% and user request response time reduced by 30 %.

Scenario 2: Two Datacentre with 25 VMs

As the application grows in popularity, it is then deployed by the provide in few more locations. This is what scenario is representing. With this assumption, two datacentres each with 25 VMs has been used in this scenario. Correspondingly, we used the "Service Proximity Based" (Closest Datacentre) and "Performance Optimized Routing" service broker policies. Figure 6 and 7 shows simulation results.

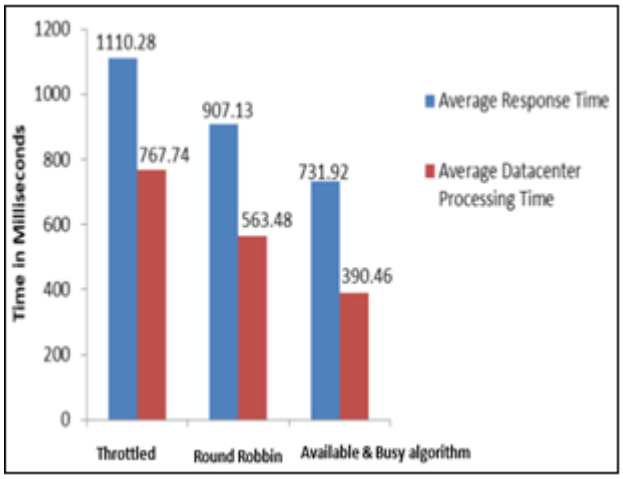

Figure 3: Two Datacentres with 25 VMs and Closest Datacentre Service Broker Policy

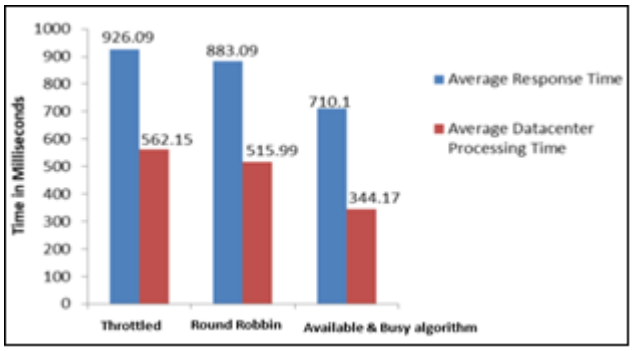

Figure 4: Two Datacentres with 25 VMs and "Performance Optimized Routing Service Broker Policy"

Licensed Under Creative Commons Attribution CC BY

Above Figure 6 and 7; clearly shows that the "Performance Optimized Routing" service broker policy offers better results. The motive behind this is that the nearest datacentre service broker policy chooses the closest datacentre by not considering the response time while the performance optimized service broker policy calculates an estimation of response time from each datacentre and chooses the best datacenter [1]. From the two cases, the proposed "BUSY & AVAILABLE "Algorithm has reduced datacentre processing time by 30% and response time magnificently by 20%.

6. Conclusions and Future Work

One of the major factors considered to improve performance of cloud computing challenges is load balancing. The primary objective of the proposed "BUSY and AVAILABLE" is to enhance response time and data processing time. In the algorithm average capacity of all virtual machines and current capacity of each virtual machine is calculated. The proposed algorithm allocates the requests to a suitable virtual machine basing both capabilities and predefined threshold values so as to avoid overloading. The results carried in a heterogenous cloud environment proved a magnificent reduction in the two considered performance parameters compared to "Throttled" and "Round Robin". In future studies it will be more imperative to compare load balancing algorithms on other parameters not just the two used in this research.

References

[1] Badshaha P Mulla, C. Rama Krishna and Raj Kumar Tickoo, "LOAD BALANCING ALGORITHM FOR EFFICIENT VM ALLOCATION IN HETEROGENEOUS CLOUD, " *International Journal of Computer Networks & Communications (IJCNC),* vol. Vol.12, no.1, pp.83 - 96, 2020.

[2] Goraya, A. Thakur and M. S., ""A Taxonomic Survey on Load Balancing in Cloud", " *Journal of Network and Computer Applications,* vol.98, pp.43 - 57, 2017.

[3] Li Daming, Su Qinglang, Deng Lianbing, Cai Kaicheng, Cai Zhiming, Bayan Omar Mohammed, "Load balancing mechanism in the cloud environment using preference alignments and an optimisation algorithm, " *The Institution of Engineering and Technology,* vol.14, no.3, pp.489 - 496, 2020.

[4] Rahmani, M. Mesbahi and A. Masoud, "Load Balancing in Cloud Computing: A State of the Art Survey, " *International. Journal of Modern Education and Computer Science,* vol.8, no.3, pp.64 - 78, 2016.

[5] Worldwide, S., "All You Need to Know About Load Balancing, " 27 May 2020. [Online]. Available: https: //www.tek - tools. com/apm/load - balancing - tools. [Accessed 6 november 2021].

[6] Ojha, Rajkumar Somani and Jyotsana, "A Hybrid Approach for VM Load Balancing in Cloud using Cloudsim, " *International Journal of Science, Engineering and Technology Research,* vol.3, no.6, pp.1734 - 1739, 2014.

[7] Nilesh, Acharya Mitali, "LOAD BALANCING IN CLOUD COMPUTING, " *International Journal of Computer Engineering & Technology (IJCET),* vol.8, no.6, pp.54 - 59, 2017.

[8] Sharma, S Sharma and S Singh and M, "Performance Analysis of Load Balancing Algorithms, " *World Academy of Science, Engineering and Technology,* vol.38, no.3, pp.269 - 272, 2008.

[9] Bharat Khatavkar, Prabadevi Boopathy, "Efficient WMaxMin Static Algorithm For Load balancing in cloud computation, " in *International Conference on Innovations in Power and Advanced Computing Technologies*, Vellore, India, 2017.

[10] N. X Phi, C. T Tin, K Thu, and T. C Hung, "Proposed Load Balancing Algorithm to Reduce Response Time and Processing Time on Cloud Computing, " *International Journal of Computer Networks and Communication,* vol.10, no.3, pp.87 - 98, 2018.

[11] Keith R Jackson, LavanyaRamakrishnan, Krishna Muriki, Shane Canon, ShreyasCholia, John Shalf, Harvey J Wasserman, and Nicholas J Wright, "Performance Analysis of High Performance Computing Applications on the Amazon Web Services Cloud, " *IEEE second international conference on cloud computing technology and science,* pp.158 - 168, 2010.

[12] Arif, AtyafDhari and Khaldun I, "An Efficient Load Balancing Scheme for Cloud Computing, " *Indian Journal of Science and Technology,* vol.10, no.11, pp.1 - 8, 2017.

[13] miniwatts, "Internet world stats, " miniwatts marketing group, 03 july 2021. [Online]. Available: https: //www.internetworldstats. com/facebook. htm. [Accessed 10 November 2021].

[14] J. A. Ashalatha R, ""Evaluation of Auto Scaling and Load Balancing Features in Cloud, ", " *International Journal of Computer Applications (0975 – 8887),* vol. vol.117, no.6, pp.30 - 33, May 2015.

[15] affiliates, Oracle Corporation and/or its, "Sun Java System Application Server Enterprise Edition 8.2 Deployment Planning Guide, " Oracle, 2010. [Online]. Available: https: //docs. oracle. com/cd/E19900 - 01/819 - 4741/abfch/index. html. [Accessed 10 November 2021].

[16] Goraya, A. Thakur and M. S., "A Taxonomic Survey on Load Balancing in Cloud, " *Journal of Network and Computer Applications,* vol.98, pp.43 - 57, 2017.

[17] Jaiswal, Ratan Mishra and Anant, "Ant colony Optimization: A Solution of Load, " *International Journal of Web & Semantic Technology (IJWesT),* vol.3, no.2, pp.33 - 50, 2012.

[18] Grance, P. Mell and T., ""The NIST Definition of Cloud Computing Recommendations of the National Institute of Standards and Technology",, " Nist Special Publication, 2011.

Licensed Under Creative Commons Attribution CC BY